RSAC 2025 Conference

Many Voices.
One Community.

SESSION ID: PART3-W09

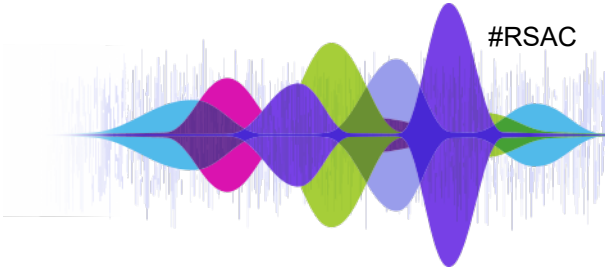
AI for SOC teams: What's real and what's next

James Spiteri

Director of Product Management
Elastic Security
linkedin.com/in/jamesspiteri/

Mike Nichols

Vice President of Product Management
Elastic Security
linkedin.com/in/howardnicholslinkedin/



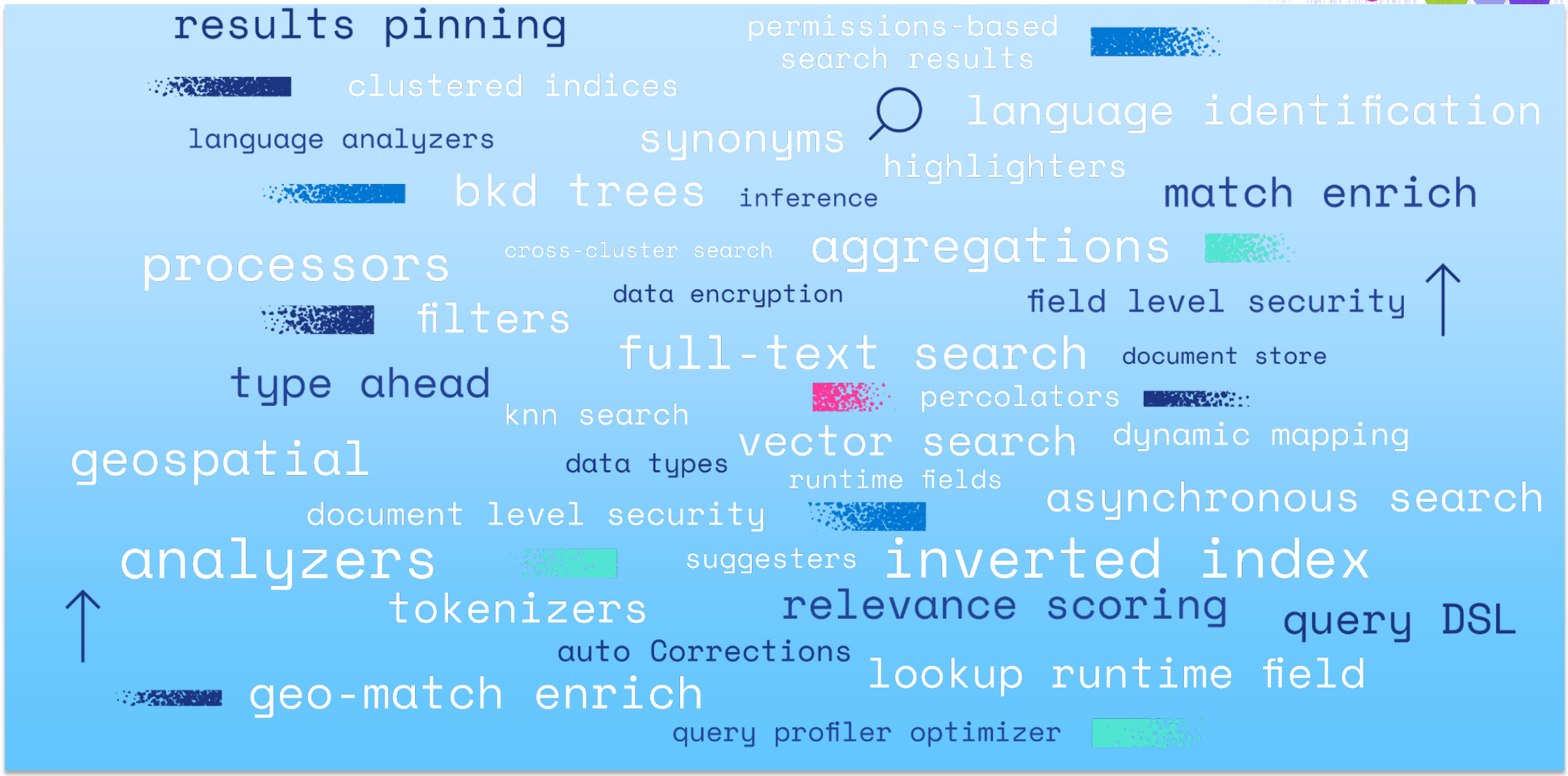
Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

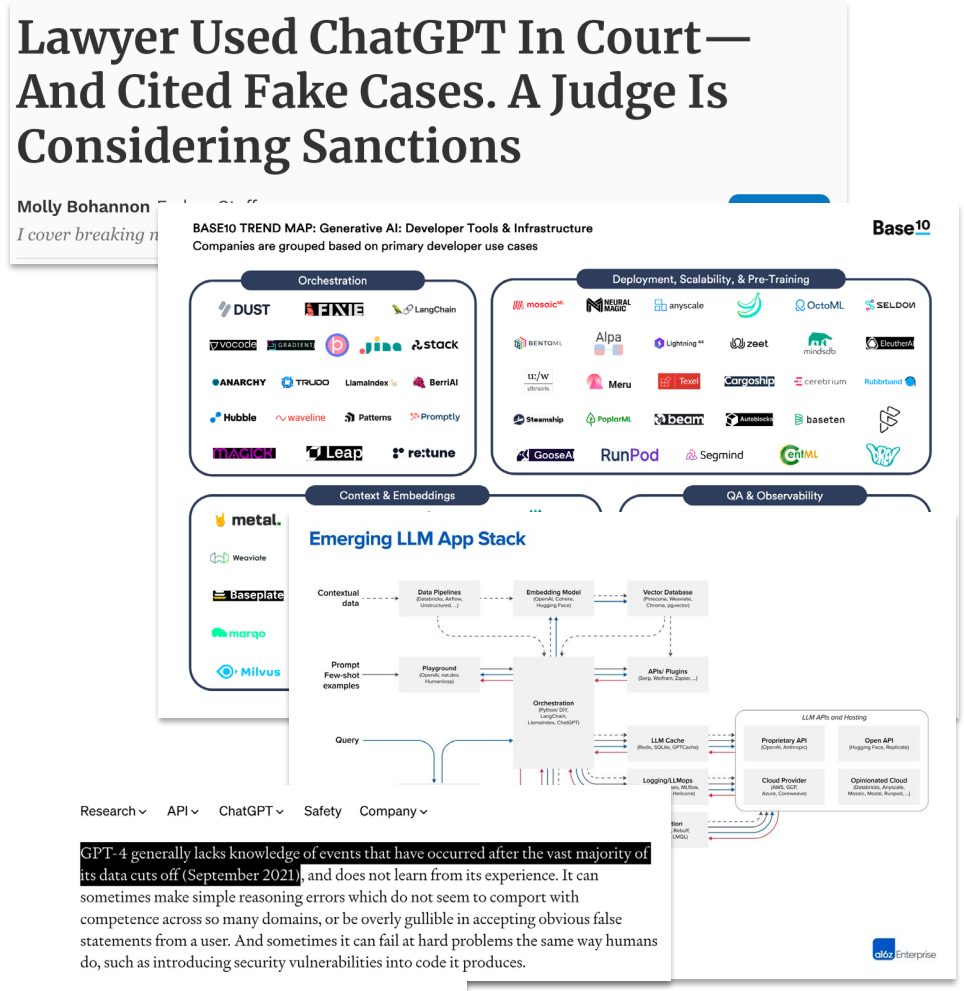
© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

We've come a long way



Generative AI comes with specific challenges

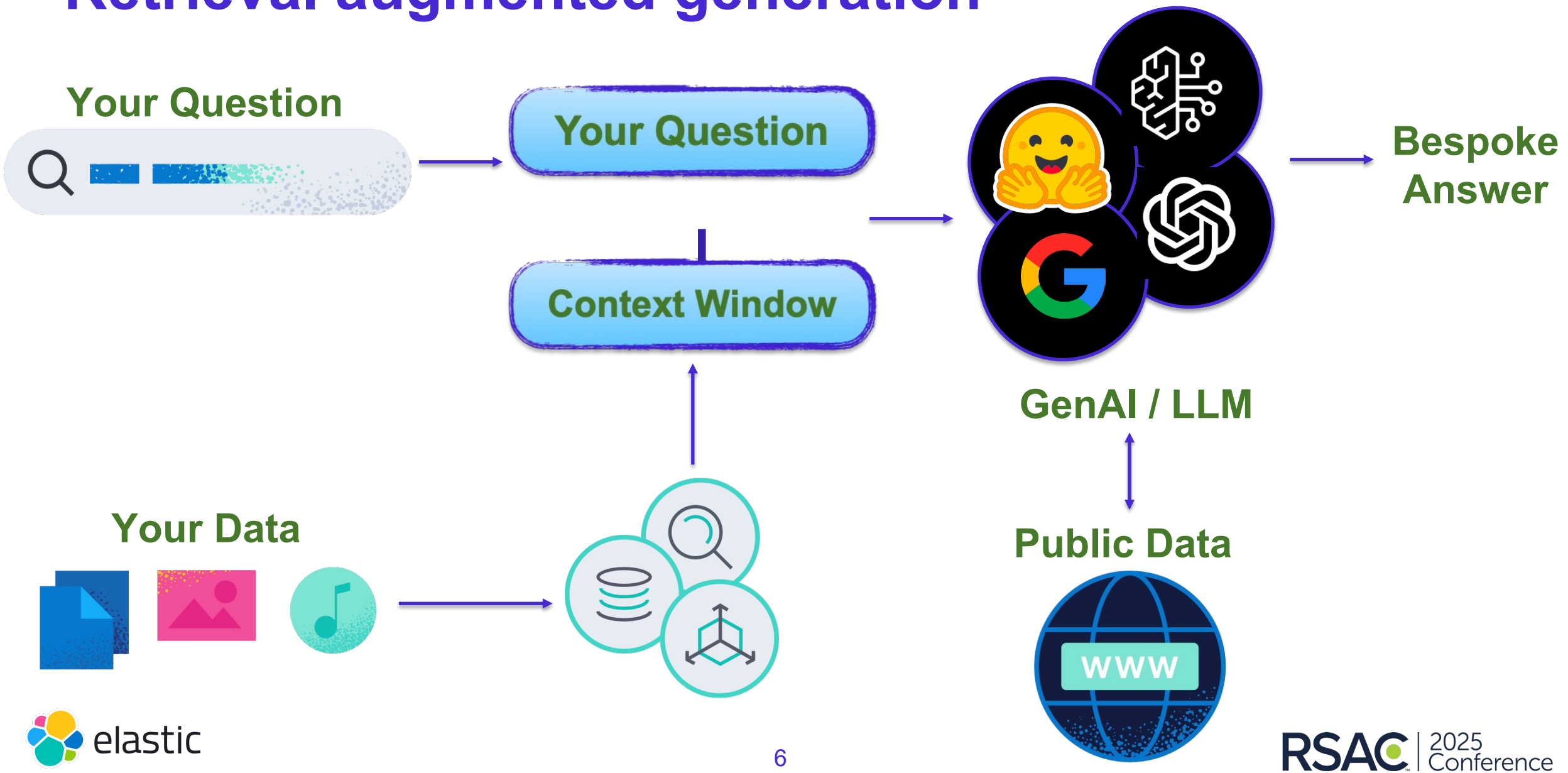
- Hallucination and bad answers
- Complex tech stack
- Real-time access to private data
- Security and privacy
- …


Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions
Molly Bohannon
BASE10 TREND MAP: Generative AI: Developer Tools & Infrastructure
Companies are grouped based on primary developer use cases
Base10
Orchestration
Deployment, Scalability, & Pre-Training
Context & Embeddings
QA & Observability
Emerging LLM App Stack
OpenAI
Research
API
ChatGPT
Safety
Company
GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its data cuts off (September 2021), and does not learn from its experience. It can sometimes make simple reasoning errors which do not seem to comport with competence across so many domains, or be overly gullible in accepting obvious false statements from a user. And sometimes it can fail at hard problems the same way humans do, such as introducing security vulnerabilities into code it produces.
Samsung bans use of generative AI tools like ChatGPT after April internal data leak
Kate Park @kateparknews / 9:17 AM EDT • May 2, 2023
Comment

Building with GenAI: Spectrum of approaches

	Pre-training	Fine-tuning	Retrieval Augmented Generation
Objective	Next-token prediction	Increase task performance	Increase task performance for specific set of domain documents
Training duration	Days to weeks to months	Minutes to hours	Not required
Expertise	High	Medium	Low
Time to adapt to new data	High; must fully retrain model	Medium; tune existing model with new information	None; new information is passed by retrieval (search)
Data requirements	Requires large amount of domain training data	Add domain-specific data	Run at inference, query specific data retrieval
Risk of hallucinations	High	High	Low
Customisation	- Can customise model architecture, size, tokenizer etc. - Creates new "foundation" LLM model	- Tune LLM for specific tasks. - Updates LLM model	- External information retrieval system can be tuned to align with LLM - Prompt can be optimised for task performance

Retrieval augmented generation



What do we need to be able to apply RAG techniques?

- Embeddings model
- Vector + data store
- Vector + BM25 search engine
- LLM
- LLM orchestration framework or functions

What do we need to be able to apply RAG techniques?

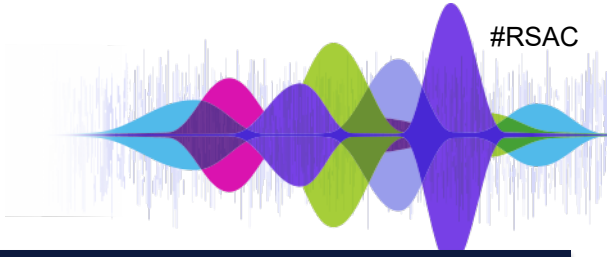
- Embeddings model => Creates embeddings/vectors
- Vector + data store => Stores data and generated embeddings
- Vector + BM25 search engine => Searches vectors and data traditionally
- LLM => Formulates the response, identifies what data to query
- LLM orchestration framework or functions => Stitches the process together

```
"text": [
  "Its use in post-exploitation \nactivities demonstrates the power of its capabilities.
   \n\nSightings of Sliver increased significantly from \nlast year. DONUTLOADER amounted to
   6.62% \nof infections, and serves as a loader to execute \nadditional malicious payloads
   in memory, avoiding \ndisk-based detection mechanisms.\n\nhttps://malpedia.caad.fkie.
   fraunhofer.de/details/win.cobalt_strike\nhttps://www.elastic.co/security-labs/
   unmasking-financial-services-intrusion-ref0657\nhttps://malpedia.caad.fkie.fraunhofer.de/
   details/win.meterpreter\nhttps://www.elastic.co/security-labs/
   linux-detection-engineering-with-auditd\nhttps://malpedia.caad.fkie.fraunhofer.de/details/
   win.sliver\nhttps://malpedia.caad.fkie.fraunhofer.de/details/win.
   donut_injector\n\n\nMalware Detections3\n\n2024 Elastic Global Threat Report\n\n10\n\nThe
   distribution of malware families reiterates the \nimportance of endpoint instrumentation
   capable of \nidentifying and mitigating malicious software, and \nenterprises that rely on
   visibility over capability may \nexperience worse outcomes than those who don't. \nThe use
   of OSTs by threats is a strong indicator \nthat technical innovations, procedural
   maturity, \nadoption of least privilege, and information sharing \nis impacting malicious
   actors. \n\n\n\nEndpoint Behaviors4\n\n2024 Elastic Global Threat Report\n\n11\n\nSecurity
   practitioners often define a threat in terms of tactics, techniques, and procedures
   (TTPs). \nIn other words, what are the goal-aligned methods used by threats? This section
   describes the most \ncommonly observed tactics, techniques, and procedures (TTPs) on
   Windows, macOS, and Linux.\n\nEndpoint\nBehaviors\n\nDistribution by operating
   system\n\nFigure 3: Endpoint behavior alerts by operating system\n\nLinux 3.30% macOS 3.
   97%\n\nWindows 92.73%\n\nWindows\nThe majority of endpoint behaviors we observed \nwere
   seen on Windows hosts, accounting for \n"
]
```



```
{
  "embeddings": [
    {
      "202": 0.45990556,
      "2019": 0.70616883,
      "2020": 0.26773226,
      "elastic": 1.9640033,
      "##eme": 1.6303077,
      "tel": 1.5551627,
      "sighting": 1.5298145,
      "tt": 1.4951544,
      "mac": 1.4771944,
      "rare": 1.4445355,
      "linux": 1.4141275,
      "population": 1.4042333,
      "##ps": 1.4017628,
      "detection": 1.3826929,
      "persistence": 1.3747752,
      "##point": 1.3458374,
      "sharing": 1.3370584,
      "threat": 1.3323604,
      "tactics": 1.3208431,
      "behavior": 1.302475,
      "populations": 1.2839587,
      "windows": 1.2772076,
      "exploit": 1.2763481,
      "host": 1.2653797,
      "##sion": 1.2243227,
      "##os": 1.203221,
      "events": 1.1956465,
      "statistics": 1.1667188,
      "##try": 1.1548659,
      "hosts": 1.1076921,
```

Evolution of the SIEM



SIEM 1.0 - 2000s

The birth of SIEM

- Log centralization
- Compliance and IR focus
- Operational features
- Visibility for the security team

Legacy

SIEM 2.0 - 2010s

Detections, Analytics and Orchestration

- Finding threats missed by others
- ML augments rules-based detection
- Analytics-driven investigations
- Response and remediation workflows

Current

SIEM 3.0 - 2023+

Generative AI Revolution

- AI driven embedded workflows
- Focus on relevance and automation
- Prioritize attacks, not alerts
- Addressing the skill gap challenge

Future

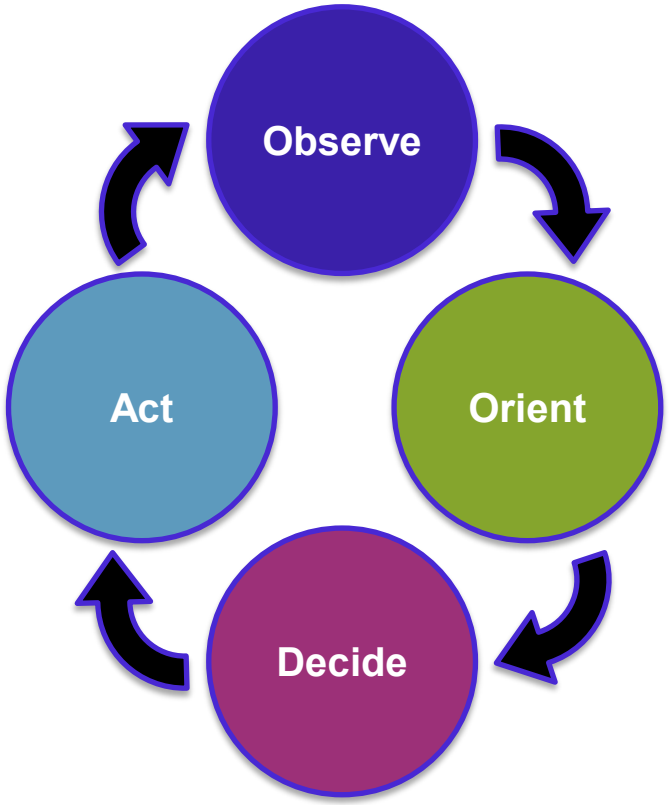
Automate and elevate your team: *Preview of the demonstration*

AI automation
- Guide analysts through triage, investigation, and response.

Natural language rule creation
- Create new detection rules using natural language with AI error detection.

AI Assistant
- Guide analysts through triage, investigation, and response.
- Get custom content and bespoke answers powered by RAG.

Natural language investigation
- Easily pivot and hunt with natural language.



Automatic Troubleshooting
- Resolve conflicts for real-time data collection and protection.

Automatic Import
- Easily onboard new data in minutes, not days.

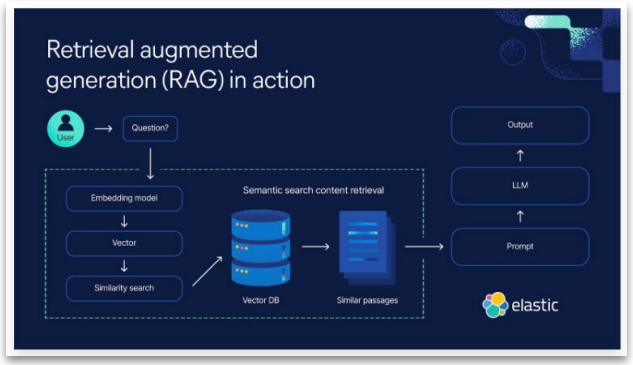
Automatic Migration
- Convert legacy SIEM detections in minutes to instantly operationalize your data.

Attack Discovery
- Prioritize attacks, not alerts; assess alerts holistically, rather than as one-off events.
- Get guidance on what to do next, ask follow-up questions, and take action.

Custom knowledge via RAG
- Bring all relevant information to the analyst at the time of investigation.

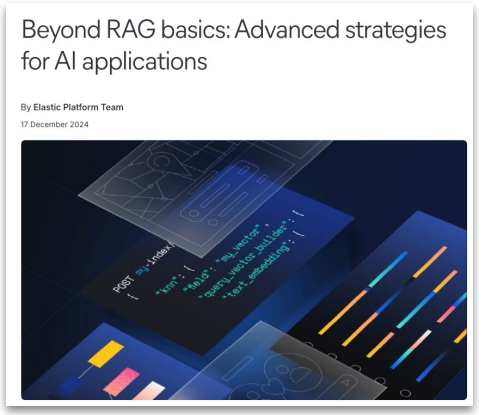
How can learn more about RAG?

- What is RAG (retrieval augmented generation)?
  - Retrieval augmented generation (RAG) is a technique that supplements text generation with information from private or proprietary data sources.



https://www.elastic.co/what-is/retrieval-augmented-generation

- Beyond RAG basics: Advanced strategies for AI Applications
  - Parallel queries
  - RAG with tools
  - Agentic RAG



https://www.elastic.co/blog/beyond-rag-basics

RSAC | 2025 Conference

RAG Demonstration

RSAC | 2025 Conference

Many Voices.
One Community.

elastic.co/security